Journal of
Computer & Electrical and Electronics Engineering Sciences
Original Article
DOI: 10.51271/JCEEES-0015

Diabetes prediction utilizing soft voting classifier

Erkan Akkur

Turkish Medicines and Medical Devices Agency, Ankara, Turkiye

Cite this article: Akkur, E. (2024) Diabetes prediction utilizing soft voting classifier. *J Comp Electr Electron Eng Sci*, 2(1), 31-34

Corresponding Author: Erkan Akkur, eakkur@gmail.com

Received: 02/04/2024 ◆ **Accepted**: 23/04/2024 ◆ **Published**: 27/04/2024

ABSTRACT

Diabetes is one of the dangerous diseases that bring about abnormalities in blood sugar levels. Early treatment can mitigate the negative consequences of this disease. Machine learning algorithms can be leveraged to predict this disease at an early stage. In this study, a soft voting ensemble classifier approach combining random forest, AdaBoost and gradient boost algorithms is adopted to predict diabetes with the highest possible accuracy. The proposed method was tested on a publicly available dataset. The proposed approach predicted diabetes with 100% accuracy. As a result of the experiments conducted within the scope of the study, polyuria and polydipsia variables were found to be the most significant risk factors for this disease. The suggested approach outperformed similar studies in the literature.

Keywords: Diabetes prediction, machine learning, ensemble learning, soft voting classifier

INTRODUCTION

Diabetes is a chronic disease characterized with abnormal levels of sugar in the blood. It may occur when insulin is not produced enough and cells are not sensitive enough to its action. The global statistics available indicate that this disease affected 529 people in 2021. Should the incidence of diabetes maintain its current rate, it is projected that 1.3 billion people will be affected by diabetes in 2050. This situation underscores the prevalence and prominence of this disease as a global health problem (Ong et al., 2023). Diabetes is characterized by several contributing factors and human errors that render the diagnosis of this disease complex. A blood test may not provide sufficient information for an adequate diagnosis of the disease. Generally, the initial symptoms of diabetes are so subtle that not even an experienced physician can identify them accurately (Alam et al., 2021). Diagnosing the disease at an early stage and identifying risk factors is crucial to address the increasing challenges in preventing diabetes and the barriers in managing the disease and its complications, as it has become a mandatory component of healthcare delivery worldwide.

Machine learning (ML) is a process of analyzing and mining data at a large scale (big data) to help discover knowledge. There has been substantial attention on ML and data mining approaches for the diagnosis, management and other associated clinical management of diabetes in recent years. By attaining high prediction rates, such algorithms can help physicians with accurate prognostic predictions based on patient clinical data and can be leveraged for diagnostics in new patient enrolments (Chaki et al., 2022). Furthermore, it is feasible to boost the performance of ML algorithms by implementing different methods. Ensemble learning (EL) techniques, which attempt to build models based on multiple classifiers instead of a single classifier, are one of the methods used to enhance model performance by compensating for the disadvantages of single classifiers. When multiple classifiers are applied together to train the input data, the actual predictions may outperform the result obtained by a single classifier (Mienye et al., 2022).

This study aims to predict diabetes disease and identify risk factors using a soft-voting ensemble learning model by leveraging a dataset generated following the risk factors of diabetes disease.

LITERATURE SURVEY

There are many studies in the literature on self-management, automatic detection, diagnosis and self-management of diabetes through ML algorithms. Laila et al. (2022) tested different ensemble ML algorithms to predict risk factors in the early stage of diabetes disease. Experiments were performed on a dataset collected from the UCI repository of several datasets comprising 17 risk factors. The random forest (RF) algorithm attained the highest prediction rate with 97% accuracy. Dutta et al. (2022) sought to predict diabetes at an early stage by

This work is licensed under a Creative Commons Attribution 4.0 International License.

proposing a model that addresses ensemble learning. The study leveraged methods such as missing value imputation, feature selection and k-fold cross-validation to ready the dataset for classification. The prediction model attained an accuracy of 73.5% and an AUC of 0.832. Rahman et al. (2023) suggested a ML-based approach for the prediction of this disease using socio-demographic attributes. RF algorithm yielded a higher prediction rate with 99.36% accuracy in comparison to other methods. The SHAP analysis was used to identify variables that are associated with diabetes risk. Al-Haija et al. (2022) conducted a study including a comparative analysis of various classifiers to analyse the risk factors of the disease. The study used a dataset with different symptoms, known as Diabetes Risk Prediction. The Shallow neural network (SNN) method attained an accuracy rate of 99.23%. Sen et al. (2023) attempted to predict this disease using decision tree-based ensemble learning models. The Extra Tree method yielded the highest prediction rate with 99.2% accuracy. Gundogdu (2023) introduced a model that blends the XGBoost algorithm and RF feature selection. As a result of the study, the suggested method attained an accuracy of 99.2% and an AUC of 0.993. Bhat et al. (2023) tried to predict diabetes disease using different ML algorithms. The Pima Indian diabetes dataset of 768 patients from the UCI was employed. As a consequence of the experiments, the decision tree technique performed the best with an accuracy rate of 91%.

PROPOSED METHODOLOGY

In this study, a soft voting ensemble classifier is used to classify diabetes as positive and negative. Figure 1 presents the proposed ensemble architecture.

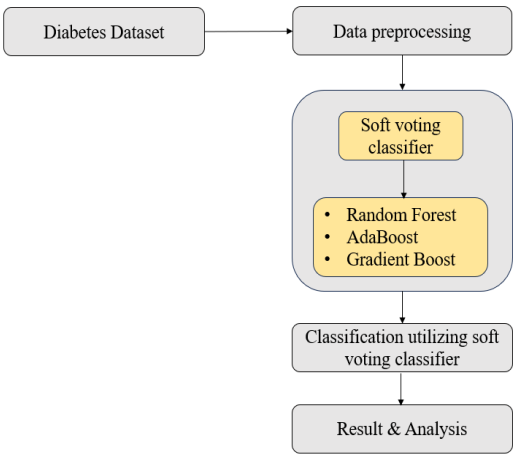

Figure 1. Suggested model architecture

Dataset

The dataset employed in this study was produced by Islam et al. (2020) and can be retrieved from the UCI machine learning repository. It is made up of data collected from 520 patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh through a questionnaire survey. Of the 520 patients, 320 were positive and 200 were negative. The dataset is organized into 16 attributes including symptoms related to diabetes. All features except age have categorical values such as Yes/No. These features are tabulated in Table 1.

Table 1. Features information of the dataset

No	Features	Range
1	age	[20-65]
2	gender	Male, Female
3	polyuria	yes, no
4	polydipsia	yes, no
5	sudden weight loss	yes, no
6	weakness	yes, no
7	polyphagia	yes, no
8	genital trush	yes, no
9	visual blurring	yes, no
10	itching	yes, no
11	irritability	yes, no
12	delayed healing	yes, no
13	partial paresis	yes, no
14	muscle stiffness	yes, no
15	alopecia	yes, no
16	obesity	yes, no

Data pre-processing

Data pre-processing is a crucial step in converting data into a handy and productive format that can be ingested into a ML algorithm (Garcia et al., 2016). Firstly, it was checked whether there was any missing data in the data set, and it was seen that there was no missing data in the data set. Then, the yes/no categorical values in the data set were converted into 0 and 1 numerical values by one-hot encoding method.

Model Architecture

The suggested method involves a soft voting ensemble classifier including random forest (RF), gradient boost (GB) and AdaBoost algorithms.

- **Random Forest (RF):** This model seeks to enhance the classification value in the classification process by creating more than one decision tree. In the model, the highest-scoring decision tree is selected among the independently considered decision trees (Breiman, 2001).
- **AdaBoost:** This model is a boosting algorithm that aims to obtain stronger learners through progressive fusion using multiple weak learners. The algorithm works by iteratively training weak learners and assigning higher importance to erroneous cases resulting from previous classifiers (Sevinc, 2022).
- **Gradient boost (GB):** This model adopts the gradient boosting technique to transform weak learners into strong learners. Each new decision tree created in the algorithm is based on the principle of minimizing the errors calculated in the previous tree. In the algorithm, a prediction is initially derived with the generated decision tree. The difference between the prediction and the target is calculated. In each new iteration, a new tree is formed with the calculated difference. As a result, the aim is to zero the difference between the prediction and the target (Aziz et al., 2020).
- **Soft voting ensemble classifier (SVE):** The EL algorithms are methods that aim to bring together different classifiers called individual learners and can provide successful results in predictive studies. The SVE method is a flexible, easy and powerful EL approach that can yield high performance in classification problems. It classifies the input data according to the probability of all predictions generated by the different individual classifiers. This method seeks to sum the prediction probabilities produced by the individual models for the class labels and to predict the class label with the highest probability (Ruta et al., 2005).

RESULTS AND DISCUSSION

The SVE model combining AdaBoost, RF and GB algorithms was utilized to predict diabetes risk. For the training and testing process, the dataset was randomly subdivided into 80% training data and 20% test data. To determine the optimum hyperparameters of the ML algorithms, hyperparameter tuning was conducted with the GridSearchCV procedure in the Sklearn library (Pedregosa et al., 2011). Table 1 illustrates the best parameter combination resulting from the grid search.

Accuracy, precision, recall and F1-score performance metrics were adopted to evaluate the robustness and efficiency of the algorithms, respectively. Table 2 presents the performance of the individual classifiers for the prediction of diabetes. When the results were analyzed, the RF algorithm attained a better prediction rate than the other models with 97.6% accuracy, 98.45% precision, 96.88% recall and 97.66% F1-Score.

Table 2. The best hyperparameters of ML algorithms

RF	AdaBoost	GB
n_estimators=200	n_estimators=100	n_estimators=200
max_depth=50	learning_rate=0.01	learning_rate=0.1
random_state=42	random_state=42	random_state=42

The SVE approach was utilized to improve the performance of the individual classifiers and provide an efficient prediction rate. Figure 2 shows the confusion matrix of the proposed model for correctly or incorrectly predicted diabetes. The SVE approach correctly classified all positive and negative examples in the dataset.

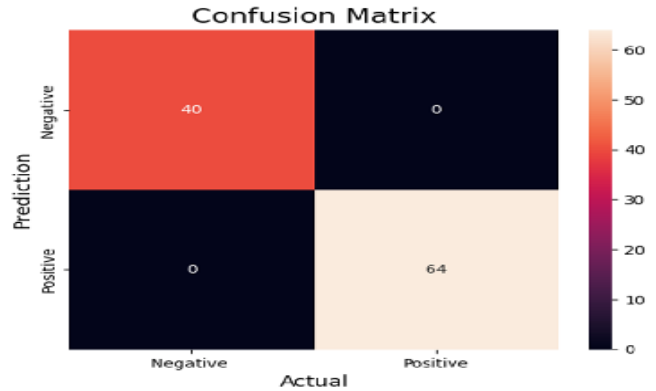

Figure 2. The confusion matrix of suggested soft voting classifier

The comparative analysis graph of the individual classifiers and the proposed SVE approach to diabetes disease prediction is presented in Figure 3. The SVE technique enhanced the performance of the individual classifiers and attained 100% accuracy.

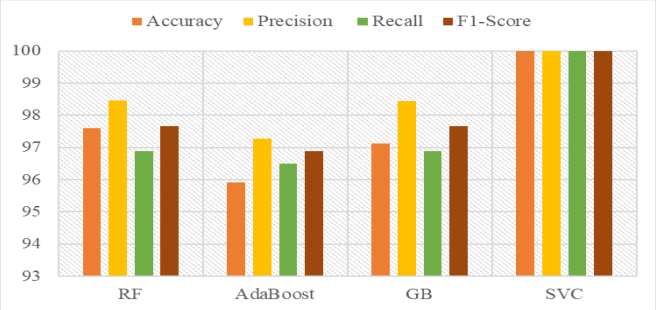

Figure 3. The comparison of individual classifiers with the suggested soft voting classifier approach

Table 3. Prediction results of the models according to performance measures

Models	Accuracy	Precision	Recall	F1-score
RF	97.6	98.45	96.88	97.66
AdaBoost	95.92	97.28	96.49	96.88
GB	97.12	98.44	96.88	97.66

In addition to classification, ML algorithms can measure the relative importance of each feature in a dataset. Figure 4 depicts the feature relative scores of RF, AdaBoost and GB algorithms for the diabetes dataset. Accordingly, all three algorithms considered polyuria and polydipsia considered as the most significant risk factors for diabetes mellitus. Polyuria is defined as excessive urine secretion and polydipsia as excessive thirst. The studies in the literature on the risk factors of diabetes mellitus indicate that polyuria and polydipsia are the most prominent risk factors as reported in this study (Pawar et al., 2017; Sekowski, 2022). The presence of polydipsia and polyuria may indicate elevated blood sugar levels in the body. It is vital to control blood sugar to prevent any health complications that may arise from this condition (Balaji, 2019).

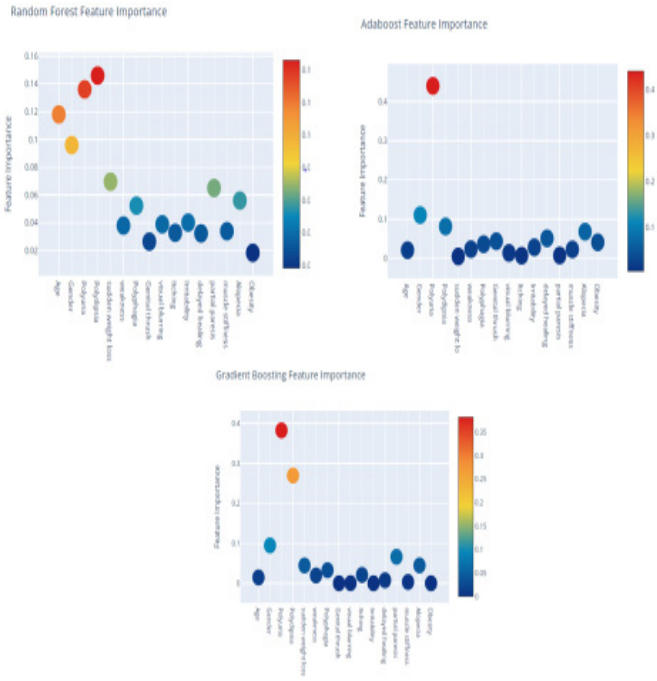

Figure 4. The feature relative scores of diabetes

Table 4 shows the classification performance of the proposed prediction model for the prediction of diabetes compared to the studies conducted in the literature using the same dataset. After the comparisons, the proposed model achieved a high prediction rate compared to similar studies in the literature.

Table 4. Comparison of the suggested model with similar studies in the literature

Study	Model	Accuracy (%)
Laila et al. (2022)	RF	97.00
Sen et al. (2023)	Extra Tree	99.20
Gundogdu (2023)	XGBoost	99.20
Rahman et al. (2023)	RF	99.36
Suggested Model	SVE	100

CONCLUSION

Diabetes can affect many people dangerously today. Early diagnosis can mitigate the consequences of this disease. The method proposed in this study has achieved remarkable results in predicting diabetes. In addition, the variables of polyuria and polydipsia, which are selected to be the most significant risk factors of diabetes, are consistent with the studies in the literature. Investigating the effectiveness of deep learning algorithms for diabetes prediction is planned in future studies.

ETHICAL DECLARATIONS

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Alam, S., Hasan, M. K., Neaz, S., Hussain, N., Hossain, M. F., & Rahman, T. (2021). Diabetes mellitus: insights from epidemiology, biochemistry, risk factors, diagnosis, complications and comprehensive management. *Diabetology,* 2(2), 36-50.
2. Al-Haija, Q. A., Smadi, M., & Al-Bataineh, O. M. (2021). Early stage diabetes risk prediction via machine learning. In *International Conference on Soft Computing and Pattern Recognition* (pp. 451-461). Cham: Springer International Publishing.
3. Aziz, N., Akhir, E. A. P., Aziz, I. A., Jaafar, J., Hasan, M. H., & Abas, A. N. C. (2020, October). A study on gradient boosting algorithms for development of AI monitoring and prediction systems. In *2020 International Conference on Computational Intelligence (ICCI)* (pp. 11-16). IEEE.
4. Balaji, R., Duraisamy, R., & Kumar, M. P. (2019). Complications of diabetes mellitus: A review. *Drug Invention Today*, 12(1), 98.
5. Bhat, S. S., Banu, M., Ansari, G. A., & Selvam, V. (2023). A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. *Healthcare Analytics*, 4, 100273.
6. Breiman, L. (2001). Random Forests, *Machine Learning,* 45(1), 5-32.
7. Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2022). Machine learning and artificial intelligence-based diabetes mellitus detection and self-management: a systematic review. *J King Saud University-Computer Informat Sci*, 34(6), 3204-3225.
8. Dutta, A., Hasan, M. K., Ahmad, M., et al. (2022). Early prediction of diabetes using an ensemble of machine learning models. *Int J Environ Res Public Health*. 19(19), 12378.
9. García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1), 9.
10. Gündoğdu, S. (2023). Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. *Multimed Tools Appl*. 82(22), 34163-34181.
11. Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer vision and machine intelligence in medical image analysis* (pp. 113-125). Springer, Singapore.
12. Laila U.E., Mahboob K, Khan AW, Khan F, & Taekeun W. (2022). An Ensemble approach to predict early-stage diabetes risk using machine learning: an empirical study. *Sensors,* 22(14), 5247.
13. Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: concepts, algorithms, applications, and prospects. *IEEE Access*, 10, 99129-99149.
14. Ong, K. L., Stafford, L. K., McLaughlin, et al. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet*, 402(10397), 203-234.
15. Pawar, S. D., Thakur, P., Radhe, B. K., Jadhav, H., Behere, V., & Pagar, V. (2017). The accuracy of polyuria, polydipsia, polyphagia, and Indian Diabetes Risk Score in adults screened for diabetes mellitus type II. *Med J Dr. DY Patil Univ*, 10(3), 263-267.
16. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: machine learning in Python. *J Machine Learning Res*, 12, 2825-2830.
17. Rahman, M. A., Abdulrazak, L. F., Ali, M. M., Mahmud, I., Ahmed, K., & Bui, F. M. (2023). Machine learning-based approach for predicting diabetes employing socio-demographic characteristics. *Algorithms*, 16(11), 503.
18. Ruta, D., & Gabrys, B. (2005). Classifier selection for majority voting. *Informat Fusion*, 6(1), 63-81.
19. Sękowski, K., Grudziąż-Sękowska, J., Pinkas, J., & Jankowski, M. (2022). Public knowledge and awareness of diabetes mellitus, its risk factors, complications, and prevention methods among adults in Poland-A 2022 nationwide cross-sectional survey. *Front Public Health*. 10, 1029358.
20. Sen, O., Bozkurt, K. S, & Keskin, K. (2023). Early-stage diabetes prediction using decision tree-based ensemble learning model. *Int Adv Res Engineering* J, 7(1), 62-71.
21. Sevinc, E. (2022). An empowered AdaBoost algorithm implementation: a COVID-19 dataset study. *Comput Industrial Eng*. 165, 107912.